

# Modeling of an Arabic electronic dictionary in generic silhouettes for the embedded systems

Ahmed Haddad\* and Hend Ben Ghezala\*\*

ENSI University, Tunis, Tunisia

\* Email: [Ahmed.haddad@ensi.rnu.tn](mailto:Ahmed.haddad@ensi.rnu.tn)

\*\* Email: [Henda.BG@cck.rnu.tn](mailto:Henda.BG@cck.rnu.tn)

**Abstract**—The objective of this article is to present a system of automatic generation of electronic dictionaries of the classic Arabic language, developed within RIADI laboratory (unit in Monastir). This system is part of project "oreillodule": an embedded system of synthesis, translation and recognition of the Arabic word. In this article, we will present the different stages of realization, and notably the automatic generation of these dictionaries based on an original theory: Conditions of Morphemic Structure (CMS), and lexical matrixes. In this case, objectives are bound to calculation time constraints and clutter memory that must be reduced to install the system on machines of reduced capacities (embedded System). That's why; we designed a dictionary models allowing a strong degree of reduction and precision but little redundancy. At the modeling stage of lexicon, words are grouped according to their descriptors, which are extracted from their generic silhouettes, forming thus under-lexicons indexed by a describer. To test our approach, we applied it on a system of recognition of Arabic writing, results are exposed at the end of this article.

## I. INTRODUCTION

The objective of this article is to present a system of automatic generation of electronic dictionaries of the classic Arabic language, developed within RIADI laboratory (Unit of Monastir). This system is part of the project "oreillodule": an embedded system of synthesis, translation and recognition of Arabic words.

The context of our work concerns the integration of lexical knowledge into the analytic recognition system. Our aim is to create a modeling for the lexicon in order to reduce the clutter memory and to increase the rate of recognition in the system.

## II. AUTOMATIC GENERATION OF THE DICTIONARY

The dictionary is definitely a fundamental element for a better performance of any natural language's automatic processing that is to say in terms of coverage or precision. In the same way the grammatical category game used in the labeling has a strong impact on the quality of the system [3]. It is clear that a reduced-category system will often lead to a better success rate than that of a more detailed system [6].

## II.1 CONDITIONS OF MORPHEMIC STRUCTURES (CMS)

Arabic phonemes are bound to the combinative and the very strict sequential restrictions that are stated in the shape of CMS. These conditions are rules that govern the generation of Arabic words: a word that violates a condition is not part of the Arabic language [5].

### Theoretical framework

Let  $x$  be the set of possible features defined by the linguistic theory.

Let  $C$  be a set of 28 Arabic consonants. Let  $C1C2C3$  be a triple root, with  $C1, C2$  and  $C3 \in C$ . Let  $PM[j][k]$  be the phonological matrix (with  $1 \leq j \leq 14$  and  $1 \leq k \leq 28$ ). This matrix represents the set of features of Arabic consonants. Let  $V$  be the set of the 6 Arabic vowels. Let  $C1V1C2V2C3V3$  be a triple root of vowels, with  $V1, V2$  and  $V3 \in V$ . Let  $PMv[j][k]$  be the phonological matrix of vowels ( $1 \leq j \leq 14$ : the set of features of Arabic vowels and  $1 \leq k \leq 6$ ).

Linguists have enumerated five CMS's that govern the Arabic word formation. These conditions are classified in two types: the combinative and the sequential restrictions.

### ■ COMBINATIVE RESTRICTIONS

These restrictions govern specifications of features corresponding to Arabic phonemes. In this case three rules are to state:

#### 1. CMS1: all phonemes are [-aspirate]

Any Arabic phoneme is a column of  $x$  specifications corresponding to these  $x$  features, the ( $x$ -fourteen) specifications that are not represented follow automatically from the fourteen present according to the specific conditions of classic Arabic. The condition CMS1 distinguishes the classic Arabic from numerous natural languages that oppose aspirate and non aspirate phonemes. It is the existence of such valid restrictions for all phonemes of classic Arabic that allowed the display of only fourteen features [5], among  $x$  definite possible features defined by the linguistic theory.

If  $ci \in C$  and  $ci \subset C1C2C3$  (with  $1 \leq i \leq 28$ ) (1)  
then  $MP[aspiré][i] = [0]$ .

#### 2. CMS2: all vocalic phonemes are [-nasal]

The condition CMS2 excludes the nasal vowels from the inventory of classic Arabic phonemes

If  $v_i \in V$  and  $v_i \subset C1V1C2V2C3V3$  (with  $1 \leq i \leq 6$ ) then  $MPv[nasale][i] = [0]$ .

3. CMS3: all phonemes that are [+consonantal] are [-syllabic] too

The condition CMS3 excludes [+syllabic] consonants. This rule is formulated in this way:

If  $MP[consonantal][i] = [-]$  then  $MP[syllabic][k] = [0]$ . (3)

In addition to the combinative restrictions between values of features belonging to one same segment, there are also some sequential restrictions.

▪ SEQUENTIAL RESTRICTIONS

These are restrictions that bind specifications of features belonging to the successive segments of the classic Arabic matrix. These restrictions show that any sequence of Arabic phonemes is not a morpheme-root or a possible allomorph (combinative variant of a phoneme). For example  $مَد$  and  $كجب$  are sequences of consonants allowed by the structure of the language, but not  $خخذ$  [7].

The fact that it doesn't exist any morpheme-root of which the phonological representation is "كجب" is not the sequence of any structural constraint, it is only an accidental gap: It is about an admissible combination by the structure of the language that is absent in the lexicon. However, sequences such as "خخذ" or "ذذبذ" are not possible morpheme-roots in classic Arabic. The first violates the restriction stated by the condition CMS4 and the second violates the one that is stated by CMS5:

CMS4: the condition CMS4 excludes from the set of the possible morpheme-root in classic Arabic any sequence of phonemes formed of two identical segments, in first and in second radical consonant.

CMS5: The condition CMS5 does not allow identical consonants which are [+continuant, +voiced] in first and third radical consonants.

Since the CMS's operate only on every isolated allomorph, they take into account only the internal constraints within one same morpheme. Thus, the process of verb generation requires the use of other tools, which are the lexical matrixes [2].

III. LEXICAL MATRIXES

III.1 TRIPLE LEXICAL MATRIXES (TLM)

These are bi-dimensional matrixes which represent the position of consonants in a triple root. These matrixes are extracted from the reference "تاج العروس", with some variations in order to exploit it in this work [1]. Twenty eight TLM's correspond to Twenty eight Arabic consonants. The 28 matrixes have stemmed from an elaborated statistic carried out by [2] on the "تاج العروس" dictionary, by transforming the triple roots of the dictionary into matrixes describing roots attested by this dictionary.

These are binary matrixes  $M_i$ , with  $1 \leq i \leq n$  ( $n = 28$ : number of consonants).

$M_i[j][k]$  refers to the  $C_iC_jC_k$  roots (with  $i, j$  and  $k \in [1..28]$ ) (for example  $كجبت$ KTb), such as:

$M_i[ ][ ]$  refers to the letter that is in a first position in the  $C_iC_jC_k$  root (ك) K

$M_i[j][ ]$  refers to the letter that is in a second position in the  $C_iC_jC_k$  root (ت) T

$M_i[ ][k]$  refers to the letter that is in a third position in the  $C_iC_jC_k$  root (ب) B

We can distinguish the following cases:

If  $M_i[j][k] = 1$  then the  $C_iC_jC_k$  root is a root attested by the  $تاج العروس$  dictionary (for example  $كجبت$ )

Otherwise ( $M_i[j][k] = 0$ ) then the  $C_iC_jC_k$  root is not attested by the "تاج العروس" dictionary. (for example:  $طضد$ ).

We can schematize this representation as follows:

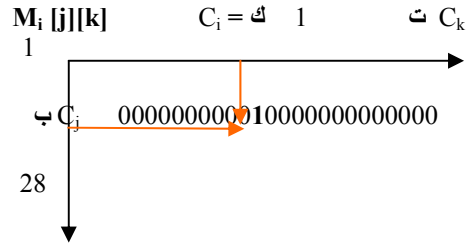


Figure 1. Representation of the lexical matrix

III.2 QUADRUPLE LEXICAL MATRIXES (QLM)

The Quadruple Lexical Matrixes are bi-dimensional matrixes that represent the position of consonants in a quadruple root. Being inspired from "تاج العروس", and "الشامل في تصريف الأفعال العربية", we have been able to establish 28 matrixes as follows:

Let  $M_i$  be a matrix, with  $1 \leq i \leq 28$ . Let  $Q$  be a representation of any quadruple root attested by Arabic. Let  $C1C2C3$  be a representation of a triple root attested and which gave the quadruple root  $Q$ , with  $C1, C2$  and  $C3 \in C$ .  $M_i[j][k]$  referring to the  $C_iC_jC_k$  roots (with  $i, j$  and  $k \in [1..28]$ ).

If  $c, d \in C$  and  $c, d \subset \{C1C2C3\}$  (such as  $c = C1$  et  $d = C2$ ) then ( $c \neq d$ ). (4)

If  $(MP[continuant][i] = [+], MP[voiced][i] = [+])$  et  $(MP[continuant][l] = [+], MP[voiced][l] = [+])$  then  $c_i, d_l \subset C1C2C3$ . (5)

These bi-dimensional matrixes are stated as follows:

- If  $M_i[j][k] = 1$  then the  $C_iC_jC_k$  root is an attested root and  $Q$  is a quadruple root generated from  $C_iC_jC_k$  by derivation with the feature "فاعل", like "كاتب".
- If  $M_i[j][k] = 2$  then the  $C_iC_jC_k$  root is an attested root and  $Q$  is a quadruple root generated from  $C_iC_jC_k$  by derivation with the feature "فعل", like "بعد".
- If  $M_i[j][k] = 3$  then the  $C_iC_jC_k$  root is an attested root and  $Q$  is a quadruple root generated from  $C_iC_jC_k$  by derivation with the feature "أفعل", like "أبعد".
- If  $M_i[j][k] = 4$  then the  $C_iC_jC_k$  root is an attested root and  $Q$  is a quadruple root  $r$  generated from  $C_iC_jC_k$  by derivation with the feature "فعلل", like "زلزل".

- If  $Mi [j][k] = x$ , with  $x \in [أ ب ج د... هـ و ي]$  then  $Q = CiCjCk x$ , like "حوقل".
- Otherwise ( $Mi [j][k] = 0$ ) then the  $CiCjCk$  root is not attested in the "تاج العروس" dictionary .

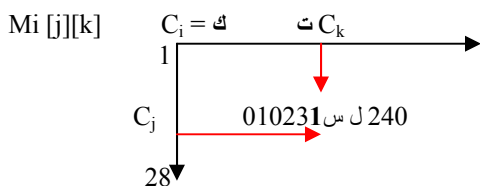


Figure 2. Representation of the quadruple lexical root

#### IV. DESCRIPTION OF THE FULFILLED SYSTEM

The automatic generation of five dictionaries of Arabic triple and quadruple roots:

The first dictionary is theoretical (21952 roots =  $(28)^3$ ). It encompasses all triple roots that are theoretically possible in standard Arabic.

The second dictionary (20415 roots): it is the dictionary of triple admissible roots.

That is to say roots that do not violate any of the C.M.S.

The third dictionary (7836): It is the dictionary of triple attested roots ; that is to say used in Arabic and taken from the table of distributions drawn from the big Arabic dictionary (الصحاح لابن الجوهري).

The Fourth dictionary (13023 roots): it is the dictionary of roots admissible by Arabic but that are not attested. These roots can be used to enrich Arabic with new words.

The fifth dictionary (4000 roots): it is the dictionary of quadruple attested roots drawn from lexical quadruple matrixes.

Certain attested triple roots do not obey one or more CSM: we created a sixth dictionary (203 roots) which gathers these roots, with for each one, the posting of the CSM which is not checked. Example: the root (بيب/bbb) is attested but does not check condition CSM4.

The capital dictionary, result of this system, contains 36876216 verbs and words of Arabic: this dictionary is generated automatically at the request of the user thus does not pose problems of obstruction in memory.

As we indicated previously, the objectives are related to the time constraints of calculation and obstruction memory which must be reduced with an aim of a bearing on machines with reduced capacities. We want to define a modeling of the dictionary allowing a strong degree of reduction but little redundancy, as well as a strong precision.

#### V. MODELING OF THE DICTIONARY:

##### V.1 FEATURES OF THE ARABIC LINGUISTICS

The Arabic alphabet is composed of 28 letters, whose shapes change according to the position in the word. The Arabic writing is semi-cursive in its two shapes, printed and handwritten. Indeed, an Arabic word is a related entity sequence entirely separated called pseudo-word. A word can be composed of one or several pseudo-words; it is due to the presence of characters that cannot be attached to their successor. Every pseudo-word is a sequence of linked letters; which gives the cursive aspect to this writing. Let's note that an isolated character can constitute a pseudo-word, as shown below:

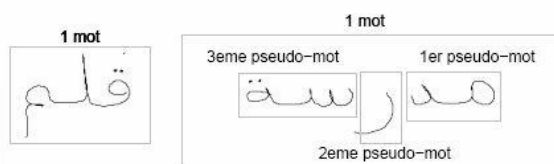


Figure 3. Example of the composition an Arabic word

Arabic writing is rich in diacritical, and especially in points. There are 15 letters, among the 28 of the alphabet that include points. These points appear above (ض), below (ي), or in the middle of the character (ج).



Figure 4. Isolated letters of the Arabic alphabet

The maximal number of diacritical that a letter can have is three points above the character or two below. These points permit to differentiate the pronunciation of Arabic letters. Therefore in a first level we propose a reduction based on the resemblance of the general shape of characters in diacritical points, then we refine our modeling according to the generic silhouette notion in order to have a more developed reduction.

##### V.2 FIRST REDUCTION LEVEL : ACCORDING TO SIMILAR FORMS

As indicated previously, the Arabic alphabet is composed of 28 letters that have some common shapes. On this resemblance rests the idea of the reduction that consists in grouping letters that have similar diacritical features in classes of shapes, then one removes the diacritical points to finally have 18 distinct shapes. Hence, without the diacritical points, the letters " Ba", " Ta " and "Tha" have the same form.

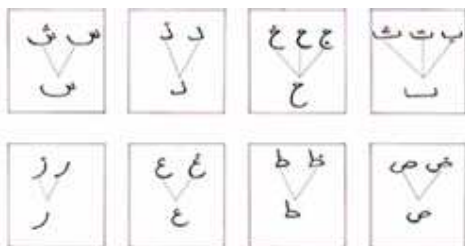


Figure 5. characters similar in form

If one assimilates letter "fa" to letter "qaf", which are only distinguished according to their position on or below the writing line, information which is not generally available unless one speaks of isolated characters or end of word characters, one has in this case 17 distinct shapes.

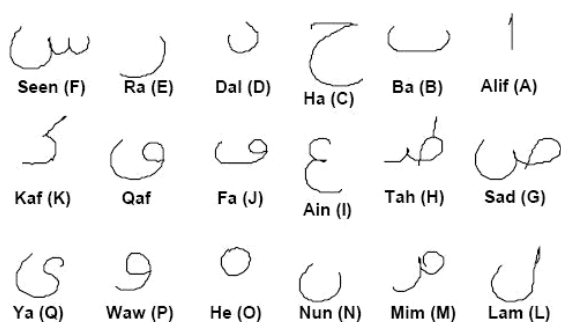


Figure 6. The 18 forms of isolated Arabic characters

This first modeling permits to create a new reduced lexicon by simplifying the global shapes of characters. It is a stage that paves the way for a more developed reduction, the modeling by generic silhouette.

### V.3 STATIC REDUCTION OF LANGUAGE : BY GENERIC SILHOUETTE

It is a specific modeling of the lexicon in view of aiming at a reduction in the time of a lexical post-treatment. The first stage consists in grouping characters in classes according to their size, their relative position in the word and the presence and the position of a vertical rod (information corresponding to the silhouette of characters), etc.

The silhouette of a character corresponds to the downward and relatively vertical fundamental features that compose it (features considered as insignificant are eliminated). The Arabic characters contain five downward features defined according to their position in the word:

- *medium* → does not exceed the body of the word,
- *upward/ hampe* → exceed upwardly,
- *downward / descendant / jambage* → exceed downwardly,
- *f-stroke* → exceed upwardly and downwardly.
- *curvature* : curves downwardly.

Label of feature	example
Medium	
Upward	
downward	
f-stroke	
curvature	

Figure 7. Fundamental features of the physical structure

The complete silhouette of a handwritten word is the concatenation of downward feature which composed. Then, it is necessary to index every word of the lexicon by an ideal describer which is its silhouette.

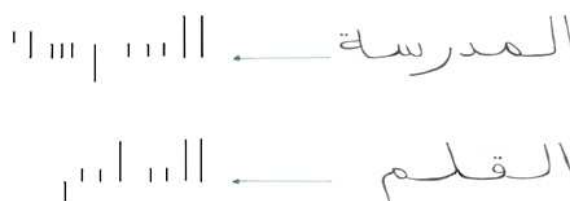


Figure 8. Example of a generic silhouette

During the lexical correction this information is taken into account and only substitutions of same class characters are accepted. The partition of the lexicon takes into account word size, its envelope (induced silhouette by silhouettes of characters) and the presence of character combinations.

As far as short words are concerned, only the information about the silhouette of characters is considered. As for the long ones, partitions are indexed by two characters and regroup all words containing these two characters. A word is placed therefore in all partitions designated by couples of characters that it contains, the number of words per partition is quite weak, but there are a lot of redundancies: on average a word is present eight times in the organized lexicon.

### V.4 MODELING OF PROPOSED LEXICON

Considering the existing means and the expected results, we opted for a static reduction of the lexicon, which is less time-consuming in calculation. The purpose is to generate the most reduced under-lexicons, while keeping a robust access criterion: the time of post-treatment decreases when the size of under-lexicons decreases, but the precision of the reduction also decreases. This reduction is characterized by:

- The 18 distinct shapes of the Arabic alphabet, which are also called when designating a word global features because of their unique aspect.
- the fundamental downward features forming the generic silhouette

We chose to use some global features for words to generate under-lexicons, these features permit us to define the

general silhouette; notion that can be determined for each word of the lexicon with the following features:

- *diacritical signs*: points above the letters « ba » ; « ta » et « tha »
- *physical structure* : composed of the most stable downward areas (features);
- *physical length*: number of fundamental downward features.

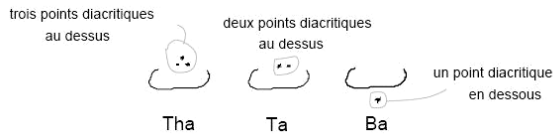


Figure 9. diacritical points of Arabic characters and their stable area

Using this complementary information aims at improving recognition, while correcting mistakes of recognition (confusion or groups of letters).

ض	ص	ض
ط	ط	ط
ظ	ظ	ظ
ع	ع	ع
غ	غ	غ
ف	ف	ف
ق	ق	ق
ك	ك	ك
ل	ل	ل
م	م	م
ن	ن	ن
و	و	و
هـ	هـ	هـ
ي	ي	ي
أ	أ	أ
ب	ب	ب
ت	ت	ت
ث	ث	ث
ج	ج	ج
ح	ح	ح
خ	خ	خ
د	د	د
ذ	ذ	ذ
ر	ر	ر
ز	ز	ز
س	س	س
ش	ش	ش
ص	ص	ص

Figure 10. Presentation of two levels of reduction

Using a representation by silhouette instead of a complete lexicon at the level of a recognition system, does not change anything at the level of the understanding but improves its performances [4].

### V.5 PRESENTATION OF THE APPROACH OF LEXICON REPRESENTATION

We present an approach of flexible lexicon modeling. The principle of this approach is: to use a simple representation of words based on global information. This representation can be defined from a handwritten word or directly on the basis of a character chain because we want it most independent possible of the writing style. During the phase of lexicon modeling, words of the lexicon are grouped according to their describers, forming thus under-lexicons indexed by a describer of under-lexicon.

During the phase of recognition, a describer of the word to recognize is extracted from its general silhouette, which

permits to select under-lexicons having the nearest describers.

### CLASSIFICATION OF WORDS BY GENERIC SILHOUETTE

The main idea of this representation is to group words having near physical structures and lengths in the same under-lexicons. The upward, downward and long features are regarded as prominent features.

In Arabic writing, the prominent features are more robust (independent of the writing style) and more easily detectable (because passing the body of the word significantly) than the median features. The physical structure of words is therefore compressed in order to be steadier and to form thus a generic silhouette: all prominent features are represented, but several successive median features are replaced by only one. According to the style of the writing (script, cursive or mixed) some words can be represented by several generic silhouettes and belong thus to several under-lexicons.

longueur -ur	silhouette générique	longueur physique	structure physique	mot manuscrit
[4-11]		8		العربية
[4-11]		9		الورقة
[4-11]		11		المدرسة
[4-11]		9		العيادة

Figure 11. Silhouette générique (les sous lexiques)

To compensate the loss of information produced by the method of compression, the generic length of the word (based on the number of downward features) is added to characterize the generic silhouette. Besides, the presence or absence of diacritical sign is taken into account. These three global features permit to organize the lexicon in under-lexicons. The visually near words (having the same generic silhouette, near generic lengths and absence or presence of diacritical signs) are regrouped in the same under-lexicons.

Le The describer of a word is represented then by a triplet (signdiacr, [lmin, lmax], S) and permits to group in an under-lexicon all words having the same generic silhouette S, of which the number of features is between lmin and lmax, and having the same value for signdiacr (0 for the absence of diacritical sign and 1 for the presence of at least one sign). A certain overlap of under-lexicons is introduced by the choice of generic length intervals to strengthen the stage of selection. Besides, some words are present in several under-lexicons or two characters having the same generic silhouette such as ص and ي.

## V.6 EXPERIMENTATIONS AND ASSESSMENTS

our objective is to integrate some lexical knowledge into a system of recognition embedded, of model the lexical treatments in order to adapt them to contexts of utilization and evolutions of the system, while taking into account the constraints of memory space and calculation time. That's why, we proposed a specific modeling of the built-in lexicon in view of a reduction. We present the different experimental results permitting to evaluate stages of the treatment. We retail the lexical resources used in these experimentations.

### ▪ LEXICON AND CORPUS

We have :

- Lexicon in the form of isolated words in Arabic, the number of words being 440,
- Corpus Corpus in the form of a text, words in a context (sentences) and the number of words being 680. Sentences have been segmented by hand to extract words. This corpus permits therefore to test the system in a context nearer to a real application.

The table below illustrates recognition rates, the term top1 is used to designate the percentage of words recognized (that is words for which the correct label is proposed at first) and top10 the percentage of words for which the correct label is among the first 10 results.

TABLE 1. RECOGNITION RATES OF RESIFMOTS WITHOUT LEXICAL KNOWLEDGE

	Base isolée	Base-contexte
Top1	39,05%	27,34%
Top10	74,34%	50,00

We have used SRILM (SRI Language Modeling Toolkit): SRILM is a box of tools for the construction and the application of language static models (LSM), it permits to identify data having the best probability and to compare them to the word to recognize.

The experimental results give birth to a significant interpretation as for the influence of the reduction of the lexicon on the performance of the handwritten recognition.

The two main measures permitting to the quality of the lexicon reduction :

- the precision of the reduction? prec red is the average number of times where the representation of the word to recognize is present in the reduced lexicon,
- the degree of reduction? deg red is linked to the average size of the reduced lexicon compared to the complete lexicon.

We compare in the table below the precision and the degree of reduction according to the modeling of the lexicon. The lexicon without organization serves as a reference.

TABLE 2. PRECISION AND DEGREE OF REDUCTION

	Isolated base		Base-contexte	
	Prec red	Deg red	Prec red	Deg red
Without modeling	100%	0%	100%	0%
Generic Silhouette	97,40%	95%	97,48%	95%

We note that the degree of reduction is less important with the modeling by generic silhouette than without modeling, it presents a value of 95%. This degree of reduction is accompanied with a decrease in the degree of precision that drops from 100% to 97,40%.

Therefore the degree of reduction and the precision of the reduction are inversely proportional; when the reduction increases, the precision decreases.

Also, rates of recognition are better when using modeling with generic silhouette than the lexicon without modeling.

This is due to the fact that modeling based on the physical structure of words introduce global information: the reduction of the lexicon guarantees that the words given as result respect the physical structure of recognition alternatives.

## VI. CONCLUSION

The context of our work concerns the integration of lexical knowledge into the analytic recognition system. Our aim is to create a modeling for the lexicon in order to reduce the clutter memory and to increase the rate of recognition in the system.

That's why we have proposed for the reduction a modeling of the lexicon leaning on a generic silhouette that permits a better degree of reduction, as well as an increase in the recognition rate.

And to test this new representation used for Arabic characters, we have resorted to the statistical tool SRILM that permits to create models of languages from the reduced lexicon and make this new model exploitable during recognition.

## REFERENCES

- [1]. A. HADDAD, (2004). Un système de génération automatique de dictionnaires linguistiques et thématiques de la langue arabe. Mastère en informatique, Ecole Nationale des Sciences de l'informatique, TUNISIE.
- [2]. A. H. MOUSSA, (1973). Statistical study of Arabic roots in mojam arous. Kouyet .

- [3]. CHANOD, J.-P and TAPANAINEN,P. (1995) : “Creating a tagset, lexicon and guesser for a french tagger ” In Proceedings of EAACL SIGDAT workshop on From Texts To Tags: Issues In Multilingual Language Analysis.
- [4]. Carbonnel. (2005) : Sabine Carbonnel ; Intégration et modélisation de connaissances linguistiques pour la reconnaissance d'écriture manuscrite en-ligne. Thesis.
- [5]. H. HABAILI, (1976). Contraintes de structure morphématique en Arabe, DEA en linguistique, Canada, université de Montréal.
- [6]. MERIALDO B, (1995):”Modèles probabilistes et étiquetage automatique” TAL, volume 36, 1995.
- [7]. T. SAIDANE, A.HADDAD, M.ZRIGUI, Pr. M. BEN AHMED, (2004). Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones JEP-TALN 2004, Fès, Maroc.